

TECHNICAL RESEARCH REPORT

Interpolation approximations for \$M|G|infty\$ arrival processes

by Konstantinos P. Tsoukatos, Armand M. Makowski

CSHCN T.R. 99-35
(ISR T.R. 99-69)



The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site <http://www.isr.umd.edu/CSHCN/>

Interpolation approximations for $M|G|\infty$ arrival processes

Konstantinos P. Tsoukatos ^{*}Armand M. Makowski [†]

Electrical Engineering Department and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract

We present an approximate analysis of a discrete-time queue with correlated arrival processes of the so-called $M|G|\infty$ type. The proposed heuristic approximations are developed around asymptotic results in the heavy and light traffic regimes. Investigation of the system behavior in light traffic quantifies the differences between the gradual $M|G|\infty$ inputs and the point arrivals of a classical $GI|GI|1$ queue. In heavy traffic, salient features are effectively captured by the exponential distribution and the Mittag-Leffler special function, under short- and long-range dependence respectively. By interpolating between the heavy and light traffic extremes we derive approximations to the queue size distribution, applicable to all traffic intensities. We examine the accuracy of these expressions and discuss possible extensions of our results in several numerical examples.

1 Introduction

The conclusions of a series of measurement studies demonstrating that network traffic exhibits persistent long term correlations have spurred recent activity in the study of queueing systems with correlated arrival processes. Analytical results reveal that, when strong dependencies are present, diverse queueing patterns may arise, in contrast to the familiar exponential decay encountered in traditional traffic models with bounded exponential moments.

^{*}The work of this author was supported through NSF Grant NSFD CDR-88-03012 and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002.

[†]The work of this author was supported partially through NSF Grant NSFD CDR-88-03012, NASA Grant NAGW277S and the Army Research Laboratory under Cooperative Agreement No. DAAL01-96-2-0002.

In this paper we deal with a discrete time queue, viewed as a surrogate for a network multiplexer, driven by an $M|G|\infty$ arrival stream. Both the discrete time $M|G|\infty$ process considered here and its continuous time variant are among the traffic models arising from large aggregations of on/off sources, that have attracted a great deal of attention. Reasons for this, such as flexibility in representing correlation functions of actual traffic traces, are discussed in [9]. A fluid queue fed by the continuous time version of the process has been studied at least as early as 1974 [3]. Later, Cox notices that the $M|G|\infty$ busy server process with heavy tailed G is a second order asymptotically self-similar process [4].

Here, we are interested in the entire steady-state queue length distribution at a multiplexer fed by the $M|G|\infty$ arrival process. For arbitrary pmf G , the system lacks the desired Markovian structure and a calculation using numerical inversion techniques, e.g. [13], is not possible, since no z -transform expressions are available. An exception is provided by the geometric case, where a two dimensional Markov chain formulation and a functional equation for the z -transform is given in [2].

To circumvent the difficulties of an exact analysis, one may rely on information gleaned from various asymptotic regimes. A promising approach consists of deriving approximations from the analysis of large buffer asymptotics [7, 8, 10]; these estimates are exact in the limit as the buffer level goes to infinity.

Our objective is to explore alternative approximations to the queue length probabilities, developed around a combination of light and heavy traffic asymptotics. Such approximations become exact in the limit as the traffic intensity goes to zero and one respectively. In light traffic we take advantage of the fact that the $M|G|\infty$ arrival process is obviously “Poisson driven”, so that the Reiman–Simon theory [12] applies, under a bounded exponential moment assumption. The resulting light traffic limits of the queue with $M|G|\infty$ arrivals differ from those of a classical $GI|GI|1$ queue. This is a manifestation of the fact that work that joins the system gradually, as is the case with $M|G|\infty$ inputs, generates less queueing than work that arrives instantaneously. From the heavy traffic regime, we collect the associated limit distribution of the queue size: This is given through the exponential function in the standard short-range dependent setup, and the Mittag–Leffler special function in the case where the $M|G|\infty$ process is long-range dependent. The approximation to the queue size distribution is subsequently generated by interpolating between the heavy and light traffic extremes. For some common pmfs G the approximant assumes a simple final form. More interestingly, it has the potential of capturing accurately the queue size distribution at small buffer sizes, for which approximations based on large buffer asymptotics are usually ill fitted. On the other hand, when G has finite exponential moment, we do not expect the heavy–light traffic interpolation to be accurate for buffer sizes much larger than the maximum burst length: It simply does not possess the correct decay rate – it does so only as the traffic intensity tends to one, i.e., in the heavy traffic limit. Surprisingly, this drawback is often absent under long-range dependence, since there are cases where the queue size distribution has hyperbolic asymptotics with the same exponent for all traffic intensities! Then an approximation is more valuable, especially when considering that, in the presence of heavy tails, alternative estimates by means of simulation take an unreasonably long time to obtain. Yet, this is

somewhat compromised by the unavailability of rigorously established light traffic limits, under long-range dependence. In Section 6 we rely on a postulated relationship, but the problem is still unresolved.

The paper is organized as follows: The description of the system model is given in Section 2. Section 3 contains the main conclusions of the light and heavy traffic analyses. These are the ingredients for the approximation, which is presented in Section 4 and discussed through numerical examples in Section 5. Further extensions of the results are suggested in Section 6.

2 The System Model

We introduce the queueing model of interest, together with the required notation. We start by presenting the $M|G|\infty$ arrival processes and several of their properties; additional facts can be found in [4, 10].

Consider a population of infinitely many information sources, operating in discrete-time. Sources can be in one of two states, active or idle. During time slot $[n, n+1)$, $n = 0, 1, \dots$, β_{n+1} new sources become active. Source j , $j = 1, \dots, \beta_{n+1}$ begins generating information by the start of slot $[n+1, n+2)$, its activity period has duration $\sigma_{n+1,j}$ (in number of slots). While active, each source emits information at a constant rate of one information unit (packet) per time slot. After its activity period expires, each source switches off permanently, never to generate packets again. Let b_n denote the number of active sources, or equivalently, the number of packets generated by the active sources at the beginning of time slot $[n, n+1)$. If initially (i.e., at time $n = 0$) there were already b active sources, we denote by $\sigma_{0,j}$ the residual activity duration (in time slots) for the j^{th} active source, $j = 1, \dots, b$.

Throughout, the \mathbb{N} -valued rvs b , $\{\beta_{n+1}, n = 0, 1, \dots\}$, $\{\sigma_{n,j}, n = 1, 2, \dots; j = 1, 2, \dots\}$ and $\{\sigma_{0,j}, j = 1, 2, \dots\}$ satisfy the following assumptions: (i) These rvs are mutually independent; (ii) The rvs $\{\beta_{n+1}, n = 0, 1, \dots\}$ are *i.i.d.* Poisson rvs with parameter $\lambda > 0$; (iii) The rvs $\{\sigma_{n,j}, n = 1, \dots; j = 1, 2, \dots\}$ are *i.i.d.* with common pmf G on $\{1, 2, \dots\}$. Let σ be a generic \mathbb{N} -valued rv distributed according to the pmf G , assume throughout that $\mathbf{E}[\sigma] < \infty$; (iv) The rvs $\{\sigma_{0,j}, j = 1, 2, \dots\}$ are *i.i.d.* \mathbb{N} -valued rvs distributed according to the *equilibrium* pmf G_e associated with G , i.e., if σ_e denotes a generic \mathbb{N} -valued rv distributed according to the pmf G_e , then

$$\mathbf{P}[\sigma_e = n] = \frac{\mathbf{P}[\sigma \geq n]}{\mathbf{E}[\sigma]}, \quad n = 1, 2, \dots \quad (1)$$

In summary, the process $\{b_n, n = 0, 1, \dots\}$ results from discrete-time Poisson(λ) arrivals of information sessions, where the session duration is distributed according to the pmf G and the packet generation rate of an ongoing session is one packet per time slot. Under the

enforced assumptions $\{b_n, n = 0, 1, \dots\}$ can be identified as the busy server process of a discrete-time $M|G|\infty$ queue; for this reason the packet arrival process $\{b_n, n = 0, 1, \dots\}$ is referred to as the $M|G|\infty$ arrival process. The following proposition shows that $\{b_n, n = 0, 1, \dots\}$ is a correlated process, with time dependencies controlled by the tail of σ [10].

Proposition 2.1 *If b is taken to be a Poisson rv with parameter $\lambda \mathbf{E}[\sigma]$, then the process $\{b_n, n = 0, 1, \dots\}$ is a (strictly) stationary ergodic process with the properties:*

- (a) *For each $n = 0, 1, \dots$, the rv b_n is a Poisson rv with parameter $\lambda \mathbf{E}[\sigma]$;*
- (b) *Its covariance function is given by*

$$\text{cov}(b_{n+j}, b_n) = \lambda \mathbf{E}[\sigma - j]^+ = \lambda \mathbf{E}[\sigma] \mathbf{P}[\sigma_e > j], \quad n, j = 0, 1, \dots$$

- (c) *Its index of dispersion of counts (IDC) is given by*

$$\text{IDC} \equiv \sum_{j=0}^{\infty} \text{cov}(b_{n+j}, b_n) = \lambda \mathbf{E}[\sigma] \sum_{j=0}^{\infty} \mathbf{P}[\sigma_e > j] = \frac{\lambda}{2} \mathbf{E}[\sigma(\sigma + 1)],$$

and the process is short-range dependent (i.e., IDC finite) if and only if $\mathbf{E}[\sigma^2]$ is finite.

We now feed this $M|G|\infty$ arrival stream $\{b_n, n = 0, 1, \dots\}$ into a discrete-time single server queue with infinite buffer capacity. Such a queueing system routinely serves as a model for a network multiplexer: If q_n denotes the number of packets remaining in the multiplexer buffer by the end of slot $[n-1, n)$, and the multiplexer output link can transmit c packets/slot, then the buffer content sequence $\{q_n, n = 0, 1, \dots\}$ evolves according to the Lindley recursion

$$q_0 = 0; \quad q_{n+1} = [q_n + b_{n+1} - c]^+, \quad n = 0, 1, \dots \quad (2)$$

From Part (a) of Proposition 2.1 the average input rate to the multiplexer is $\mathbf{E}[b_n] = \lambda \mathbf{E}[\sigma]$, and the system is *stable* if the traffic intensity $\rho \equiv \lambda \mathbf{E}[\sigma] / c$ satisfies $\rho < 1$. In that case $q_n \Rightarrow_n q$, where the \mathbb{R} -valued rv q is the stationary queue size in the multiplexer buffer. We are interested in evaluating

$$P(b, \rho) \equiv \mathbf{P}_\rho[q > b], \quad b \geq 0 \quad \text{and} \quad Q_m(\rho) \equiv \mathbf{E}_\rho[q^m], \quad 0 \leq \rho < 1, \quad m = 1, 2,$$

i.e., the probability that the stationary queue size exceeds b , and the queue size first and second moments, when the traffic intensity is ρ . To that end we develop simple approximations that all flow from asymptotic results under heavy and light traffic conditions.

3 Heavy and Light Traffic

The interpolation approximation we have in mind hinges on the availability of explicit expressions for limits of system quantities as $\rho \rightarrow 1$ (heavy traffic limits), and derivatives with respect to ρ as $\rho \rightarrow 0$ (light traffic derivatives). It thus requires examination of the behavior of the queue with $M|G|\infty$ arrivals under each one of these two asymptotic regimes.

Light Traffic We start with the light traffic regime. The right-hand derivatives at $\rho = 0$ of the various metrics of interest are evaluated using the Reiman–Simon technique [12]. For the system to be in the domain of applicability of the Reiman–Simon results, an assumption on finiteness of the exponential moment of σ is needed.

Assumption (A) *There exists $\theta^* > 0$ such that $\mathbf{E}[e^{\theta\sigma}] < \infty$ for $\theta < \theta^*$.*

The detailed light traffic analysis of the queue with $M|G|\infty$ arrivals is provided in [15]. We summarize the conclusions in the following.

Proposition 3.1 *Consider the Lindley recursion (2) with integer release rate $c = 1, 2, \dots$, and let $b = 0, 1, \dots$. Under Assumption (A) it holds that*

(a) *For each $n = 0, 1, \dots, c$*

$$\frac{\partial^n}{\partial \rho^n} P(b, 0+) = 0 \quad \text{and} \quad \frac{d^n}{d\rho^n} Q_m(0+) = 0, \quad m = 1, 2, \dots \quad (3)$$

(b) *In addition, for $c = 1$,*

$$\begin{aligned} \mathbf{E}[\sigma]^2 \frac{\partial^2}{\partial \rho^2} P(b, 0+) &= \mathbf{E}[(\sigma - b)^+]^2 \mathbf{P}[\sigma > b] + 2 \mathbf{E}[(\sigma - b)^+]^2 \\ &\quad - 3 \mathbf{E}[(\sigma - b)^+] \mathbf{P}[\sigma > b] + \mathbf{P}[\sigma > b]^2 \end{aligned} \quad (4)$$

$$\frac{d^2}{d\rho^2} Q_1(0+) = \frac{\mathbf{E}[\sigma^2]}{\mathbf{E}[\sigma]} \quad (5)$$

and

$$\frac{d^2}{d\rho^2} Q_2(0+) = \frac{1}{2} \left(1 + \frac{\mathbf{E}[\sigma^2]^2}{\mathbf{E}[\sigma]^2} \right). \quad (6)$$

Proposition (3.1) delineates a light traffic behavior for the queue with $M|G|\infty$ arrivals that is certainly different from the one of a classical $GI|GI|1$ queue. As seen from (3), when $c = 1$, in which case the multiplexer can serve no more than one source per time slot, the first derivative of the tail probability is zero. Hence, in a Taylor expansion of $\mathbf{P}_\rho[q > b]$ around $\rho = 0$ the linear term in ρ would offer no contribution. In contrast, the stationary workload W in a single server $M|G|1$ queue is known to satisfy $\mathbf{P}_\rho[W > x] \sim \rho(1 - G_e(x))$ ($\rho \rightarrow 0$), that is, in the classical queueing setup the corresponding expansion starts with a non-zero ρ term. For $M|G|\infty$ arrivals it is the second derivative (4) which is the most informative. This highlights the role of the activity duration rv σ , through both its distribution and its first two moments. Notice that even if Assumption (A) were to be relaxed, (4) shows that for $\mathbf{P}_\rho[q > b]$ to decay like ρ^2 for small ρ it is necessary that $\mathbf{E}[\sigma^2]$ be finite. If $\mathbf{E}[\sigma^2] = \infty$, as is the case for long-range dependent $M|G|\infty$ arrivals, expression (4) yields infinity and ρ^2 is no longer the correct order of decay. A different, perhaps smaller exponent should be sought

in the long-range dependent case (see (21)). Finally, relations (3) reflect (though in a rough manner) the statistical multiplexing gain: Since the first non-zero contribution to the tail probability is no lower than ρ^{c+1} , (3) implies that increasing the multiplexer release rate c while maintaining the same traffic intensity ρ would result in a decreasing tail probability $\mathbf{P}_\rho[q > b]$, as could be expected.

Heavy traffic Considered next is the behavior of the queue with $M|G|\infty$ arrivals in heavy traffic, that is, as the arrival rate $\lambda \mathbf{E}[\sigma]$ tends to the multiplexer release rate c from below. Clearly, as the traffic intensity ρ converges to one the system becomes unstable and the queue length grows unbounded. It is thus necessary to seek a suitable *normalizer* for the queue length process, so that its normalized version has a non trivial heavy traffic limit. This problem is typically addressed in a setup where the system of interest is embedded into a family of queueing systems, parametrized by an integer, say $l = 1, 2, \dots$, ensuring that, as $l \uparrow \infty$, the appropriate trend to instability is established. Such an approach was pursued in [16], providing a complete characterization of the arising heavy traffic limits. We tacitly assume here that the heavy traffic limit of the stationary distribution coincides with the stationary distribution of the heavy traffic limit, and gather the required results from [16] in a convenient form:

Proposition 3.2 *The heavy traffic limits of the stationary queue length distribution associated with (2) can be classified as follows:*

(a) *If $\mathbf{E}[\sigma^2] < \infty$ then*

$$\lim_{\rho \rightarrow 1} \mathbf{P}_\rho[(1 - \rho)q > x] = \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}[\sigma^2]}x\right), \quad x \geq 0. \quad (7)$$

(b) *If $\mathbf{P}[\sigma > n] = n^{-\alpha}$, $n = 1, 2, \dots$, with $1 < \alpha < 2$, then*

$$\lim_{\rho \rightarrow 1} \mathbf{P}_\rho[(1 - \rho)^{1/(\alpha-1)}q > x] = E_{\alpha-1}\left(-\frac{(\alpha-1)\mathbf{E}[\sigma]}{\Gamma(2-\alpha)}x^{\alpha-1}\right), \quad x \geq 0, \quad (8)$$

where

$$E_\nu(x) \equiv \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(\nu n + 1)}, \quad \nu > 0, \quad x \in \mathbb{R}, \quad (9)$$

is the Mittag-Leffler special function [5].

Part (a) of Proposition 3.2 addresses the classical short-range dependent case, for which the heavy traffic normalizer is $(1 - \rho)$ and the limiting heavy traffic distribution is exponential. Part (b) deals with a long-range dependent $M|G|\infty$ arrival process. Under long-range dependence, the heavy traffic queue length distribution is expressed through a Mittag-Leffler function with hyperbolic decay, while the power-law behavior of the heavy traffic normalizer is $(1 - \rho)^{1/(\alpha-1)}$. We mention that this result can be stated in a more general manner to cover the situation where the tail of σ is *regularly varying* of order α , $1 < \alpha < 2$.

The results under the light and heavy traffic regimes are subsequently combined into approximations for all values of the traffic intensity.

4 Interpolation Approximations

Whenever Assumption (A) is satisfied, $\mathbf{P}_\rho[q > b]$ is infinitely differentiable with respect to ρ at $\rho = 0$, hence it can be approximated by bringing together heavy traffic limits and light traffic derivatives into a Taylor series-like expansion. To this end we follow the approach proposed in [6]. In passing, we also discuss approximations for $Q_m(\rho)$, $m = 1, 2$. The details are as follows:

The interpolation method Consider the normalized queue length rv $(1 - \rho) q$ and define

$$F(x, \rho) \equiv \mathbf{P}_\rho[(1 - \rho) q > x], \quad 0 \leq \rho < 1, \quad x \geq 0 \quad (10)$$

and

$$F(x, 1) \equiv \lim_{\rho \rightarrow 1} \mathbf{P}_\rho[(1 - \rho) q > x]. \quad (11)$$

Assume that partial derivatives of $F(x, \rho)$ with respect to ρ , up to order n , at $\rho = 0+$, are available. Construct $\hat{F}_n(x, \rho)$, the n^{th} order interpolation approximation to $F(x, \rho)$, by means of the polynomial

$$\hat{F}_n(x, \rho) \equiv \sum_{i=0}^n \frac{\rho^i}{i!} \frac{\partial^i}{\partial \rho^i} F(x, 0+) + \left(F(x, 1) - \sum_{i=0}^n \frac{1}{i!} \frac{\partial^i}{\partial \rho^i} F(x, 0+) \right) \rho^{n+1}. \quad (12)$$

Observe that

$$\hat{F}_n(x, 1) = F(x, 1) \quad \text{and} \quad \frac{\partial^i}{\partial \rho^i} \hat{F}_n(x, 0+) = \frac{\partial^i}{\partial \rho^i} F(x, 0+), \quad i = 0, 1, \dots, n,$$

that is, $\hat{F}_n(x, \rho)$ is precisely that unique $n + 1$ degree polynomial in ρ which matches the $n + 1$ partial derivatives of $F(x, \rho)$ at $\rho = 0+$ and its heavy traffic limit. Now, by reversing the $(1 - \rho)$ normalization in $\hat{F}_n(x, \rho)$ we generate the n^{th} order interpolation approximation to $\mathbf{P}_\rho[q > b]$ as

$$\mathbf{P}_\rho[q > b] \approx \hat{F}_n((1 - \rho) b, \rho). \quad (13)$$

Note that, in principle, this may lie outside $[0, 1]$, in which case it is obviously a poor approximation. To calculate the quantities associated with (13) it remains to express the partial derivatives appearing in (12) in terms of the light traffic derivatives of $\mathbf{P}_\rho[q > b]$. We have

$$\frac{\partial}{\partial \rho} F(x, 0+) = \frac{\partial}{\partial \rho} P(x, 0+) + x \frac{\partial}{\partial x} P(x, 0+) \quad (14)$$

and

$$\begin{aligned} \frac{\partial^2}{\partial \rho^2} F(x, 0+) &= \frac{\partial^2}{\partial \rho^2} P(x, 0+) + 2x \frac{\partial^2}{\partial \rho \partial x} P(x, 0+) \\ &\quad + 2x \frac{\partial}{\partial x} P(x, 0+) + x^2 \frac{\partial^2}{\partial x^2} P(x, 0+). \end{aligned} \quad (15)$$

In case additional light traffic information is available, repeated application of the chain rule will yield higher order derivatives, as needed.

Approximate expressions We are now ready to write approximate expressions anchored on the heavy and light traffic results of Section 3. Proposition 3.2(a) provides the limit (11) that should be inserted in (12). Proposition 3.1(a) can be used to substitute for the partials in (14) and then in (12). Thus, if the multiplexer release rate is $c = 1, 2, \dots$, the c^{th} order interpolation approximation to $\mathbf{P}_\rho[q > b]$ is simply

$$\mathbf{P}_\rho[q > b] \approx \hat{F}_c((1 - \rho) b, \rho) = \rho^{c+1} \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}[\sigma^2]} (1 - \rho) b\right). \quad (16)$$

More can be accomplished in the case $c = 1$, since Proposition 3.1(b) affords us a promising 2^{nd} order interpolation approximation. Insertion of (15) in (12) yields

$$\hat{F}_2(b, \rho) = \frac{1}{2} \rho^2 (1 - \rho) \frac{\partial^2}{\partial \rho^2} P(b, 0+) + \rho^3 \exp\left(-\frac{2\mathbf{E}[\sigma]}{\mathbf{E}[\sigma^2]} b\right) \quad (17)$$

and the latter leads to the 2^{nd} order approximant

$$\mathbf{P}_\rho[q > b] \approx \hat{F}_2((1 - \rho) b, \rho), \quad c = 1, \quad (18)$$

where Proposition 3.1(b) is used to supply the second partial derivative in (17).

Next, we briefly deal with moment approximations. We restrict attention to the case $c = 1$ and consider only the queue length first and second moment. The relevant light traffic limits are given by (3), (5) and (6). In heavy traffic it can be inferred from (7) that

$$\lim_{\rho \rightarrow 1} (1 - \rho)^m Q_m(\rho) = m! \left(\frac{\mathbf{E}[\sigma^2]}{2\mathbf{E}[\sigma]} \right)^m, \quad m = 1, 2, \dots$$

Moment interpolations are then developed in very much the same manner as distribution interpolations. We skip the details of the derivation and list the final expressions

$$Q_1(\rho) = \frac{\mathbf{E}[\sigma^2]}{2\mathbf{E}[\sigma]} \frac{\rho^2}{1 - \rho}, \quad c = 1 \quad (19)$$

and

$$Q_2(\rho) \approx \frac{\rho^2}{4(1 - \rho)^2} \left(\rho \left(\frac{\mathbf{E}[\sigma^2]^2}{\mathbf{E}[\sigma]^2} - 1 \right) + \frac{\mathbf{E}[\sigma^2]^2}{\mathbf{E}[\sigma]^2} + 1 \right), \quad c = 1. \quad (20)$$

We stress that formula (19) is in fact *exact*. This is a result whose continuous time analog has been established for a more general fluid model in [14, p. 23]. On the contrary (20) cannot be exact. To verify this consider the example where $\sigma = 1$ deterministic. This corresponds to *i.i.d.* Poisson arrivals, for which a probability generating function of the queue length rv is available. It can be shown [15] that the exact expression is

$$Q_2(\rho) = \frac{\rho^2}{6(1-\rho)^2} (\rho^2 - \rho + 3), \quad c = 1, \quad \sigma = 1 \text{ a.s.}$$

a formula that clearly cannot be recovered using only two light traffic derivatives. Still, when $\sigma = 1$ approximation (20) is within 9% of the correct value, for all traffic intensities.

We close this section with a comment. Recall that each active source in the $M|G|\infty$ arrival process generates one information unit per time slot. So, $c = 1$ corresponds to the case where the amount of service in one slot is exactly equal to the amount of information that one active source generates in one slot. When $c = 1$ a single active source suffices to make full use of the server capacity; in this system there is never any leftover capacity to simultaneously serve more than one sources. On the contrary, when $c > 1$, the server can attend to more than one sources during one time slot, so that there is a multiple service feature to the system behavior. An exact or approximate analysis in this regime is clearly more challenging.

5 Numerical Results

To gauge the accuracy of the proposed expressions we carry out simulation experiments choosing various distributions for the burst duration rv σ . The experimental values are obtained by regenerative simulation and relative widths accompanying them correspond to 95% confidence intervals. We confine ourselves to the simple situation where the multiplexer release rate is $c = 1$. While the list of examples below is not exhaustive, it does serve to illustrate the ability of the heavy–light traffic interpolation to “ballpark” the true tail probabilities, as well as its limitations.

Deterministic When the burst duration is deterministic, $\sigma = D$ a.s., $D = 1, 2, \dots$, approximation (18) reads

$$\begin{aligned} \mathbf{P}_\rho[q > b] \approx & \frac{\rho^2(1-\rho)}{2D^2} \left(3[D - (1-\rho)b]^+ ([D - (1-\rho)b]^+ - 1) \right. \\ & \left. + \mathbf{1}[D > (1-\rho)b] \right) + \rho^3 \exp\left(-\frac{2}{D}(1-\rho)b\right), \quad b = 0, 1, \dots \end{aligned}$$

We let the burst duration be $\sigma = 3$ and obtain simulation estimates for the steady state probability $\mathbf{P}_\rho[q_\infty > 0]$. In this case an explicit expression for $\mathbf{P}_\rho[q_\infty > 0]$ is available [15].

ρ	Tail probability $\mathbf{P}[q_\infty > 0]$			
	Exact	Simulation	Approximation	Error (%)
0.1	1.0478e-02	1.0469e-02 \pm 0.2%	1.0500e-02	-0.21
0.2	4.1622e-02	4.1668e-02 \pm 0.3%	4.1778e-02	-0.38
0.3	9.3042e-02	9.3020e-02 \pm 0.2%	9.3500e-02	-0.49
0.4	1.6441e-01	1.6442e-01 \pm 0.2%	1.6533e-01	-0.56
0.5	2.5545e-01	2.5533e-01 \pm 0.2%	2.5694e-01	-0.58
0.6	3.6594e-01	3.6607e-01 \pm 0.1%	3.6800e-01	-0.56
0.7	4.9573e-01	4.9601e-01 \pm 0.1%	4.9817e-01	-0.49
0.8	6.4470e-01	6.4488e-01 \pm 0.1%	6.4711e-01	-0.37
0.9	8.1279e-01	8.1264e-01 \pm 0.1%	8.1450e-01	-0.21

Table 1: $\mathbf{P}[q_\infty > 0]$ for deterministic burst duration $\sigma = 3$.

In Table 1 we list simulation estimates, numerical values from the exact formula and from the light-heavy traffic interpolation. A comparison of the exact values to the light-heavy traffic interpolation shows that, in this case, the agreement is excellent. Since we expect the approximation to be asymptotically exact at the endpoints $\rho = 0$ and $\rho = 1$, it is not surprising that the largest errors occur in moderate traffic.

ρ	Tail probability $\mathbf{P}_\rho[q > 4]$		
	Simulation	Approximation	Error (%)
0.1	1.1271e-04 \pm 1.7%	9.0718e-05	19.51
0.2	1.3444e-03 \pm 1.7%	9.4753e-04	29.52
0.3	6.1736e-03 \pm 0.9%	5.9952e-03	2.89
0.4	1.9246e-02 \pm 0.6%	1.4415e-02	25.10
0.5	4.7745e-02 \pm 0.5%	3.9894e-02	16.44
0.6	1.0292e-01 \pm 0.4%	9.5777e-02	6.94
0.7	2.0035e-01 \pm 0.3%	1.9757e-01	1.39
0.8	3.6093e-01 \pm 0.3%	3.6379e-01	-0.79
0.9	6.1407e-01 \pm 0.3%	6.1902e-01	-0.81

Table 2: $\mathbf{P}_\rho[q > 4]$ for deterministic burst duration $\sigma = 3$.

In the same setup, we next consider the tail probability $\mathbf{P}_\rho[q > 4]$. From Table 2 we see that although the approximation yields estimates in the correct order of magnitude, the errors are substantial when not in the moderate-to-heavy traffic regime. This can be explained as follows: When $\sigma = 3$, in order for the queue to build up to 4 at least 3 sources should be simultaneously active. Note that the light traffic component of the approximation consists of the second derivative, which can be obtained by considering sample paths with at most two source arrivals in the system. Thus, any effects due to the activation of more than two sources are not adequately accounted for in light traffic.

Uniform Specializing (18) to the case where σ is uniformly distributed, $\mathbf{P}[\sigma = n] = 1/M$, $n = 1, 2, \dots, M$, yields

$$\mathbf{P}_\rho[q > b] \approx \mathbf{1}[M > (1 - \rho)b] \frac{\rho^2(1 - \rho)}{3M^2(M + 1)^2} \left(1 + 5(M - (1 - \rho)b)^2\right) \\ \times (M - (1 - \rho)b)^2 + \rho^3 \exp\left(-6 \frac{(1 - \rho)b}{2M + 1}\right), \quad b = 0, 1, \dots$$

Buffer size	Tail probability $\mathbf{P}_\rho[q > b]$		
	Simulation	Approximation	Error (%)
0	4.4907e-02±0.2%	4.5333e-02	-0.95
2	1.1747e-02±0.5%	1.1399e-02	2.97
5	1.4543e-03±1.4%	9.7380e-04	33.04
8	1.9456e-04±3.7%	2.4379e-04	-25.30
10	5.1620e-05±7.0%	1.0186e-04	-97.31

Table 3: Traffic intensity $\rho = 0.2$; $\sigma \sim \text{uniform}(1, 5)$.

For $M = 5$ we compare simulation vs approximation in Tables 3 and 4, for traffic intensities $\rho = 0.2$ and $\rho = 0.8$ respectively. Once more, the approximation is very sharp for small buffer sizes. As the buffer size increases beyond the maximum burst length and the true probabilities become smaller, the approximation lingers on in the correct order of magnitude, but it clearly deteriorates away from heavy traffic. Eventually, as the buffer size tends to infinity, the interpolation approximation overestimates the actual probabilities.

Buffer size	Tail probability $\mathbf{P}_\rho[q > b]$		
	Simulation	Approximation	Error (%)
0	6.5489e-01±0.1%	6.6133e-01	-0.98
10	2.0086e-01±0.4%	1.9161e-01	4.60
20	6.3303e-02±0.9%	5.8057e-02	8.29
30	1.9964e-02±1.7%	1.9406e-02	2.79
40	6.2827e-03±3.1%	6.5188e-03	-3.75
50	1.9703e-03±5.6%	2.1897e-03	-11.13

Table 4: Traffic intensity $\rho = 0.8$; $\sigma \sim \text{uniform}(1, 5)$.

Geometric Taking σ to follow a geometric distribution, $\mathbf{P}[\sigma > n] = p^n$, $n = 0, 1, \dots$, we obtain from (18) that

$$\mathbf{P}_\rho[q > b] \approx \frac{\rho^2}{2}(1 - \rho)(1 + p)^2 p^{2(1 - \rho)b} + \rho^3 \exp\left(-2 \frac{1 - p}{1 + p}(1 - \rho)b\right), \quad b = 0, 1, \dots$$

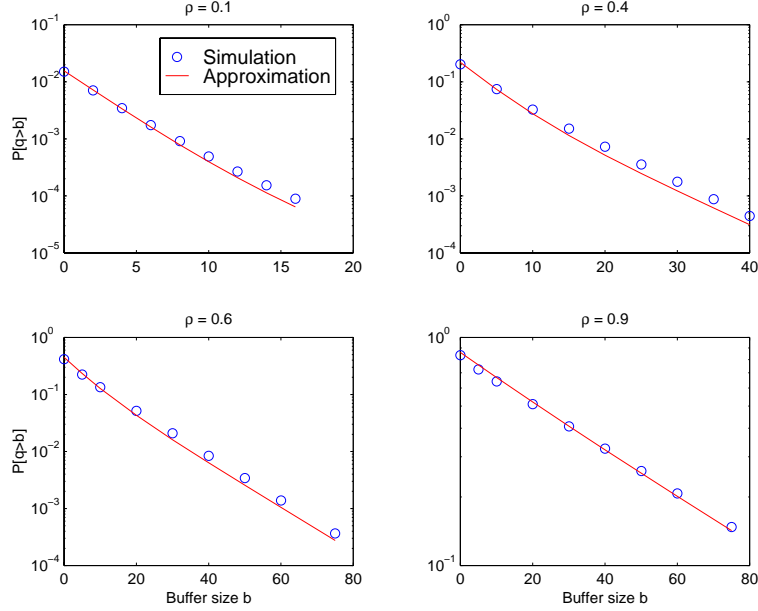


Figure 1: Geometric $p = 0.8$ burst duration.

As an example we set $p = 0.8$ and plot simulated and approximate values in Figure 1, for traffic intensities $\rho = 0.1, 0.4, 0.6$ and 0.9 . In all cases confidence interval widths were within 10% of the mean and are not shown. The linear decrease of the simulated values suggests an exponential decay of the queue length distribution, in agreement with large deviations results. Figure 1 clearly indicates that the heavy–light traffic interpolation is sufficient for providing rough estimates for a wide range of probabilities and buffer sizes.

6 Extensions

It is apparent from the developments of Section 3 that the light traffic results, as stated in Proposition (3.1), do not cover several interesting distributions belonging to the subexponential family. Such is for example the lognormal distribution, which violates Assumption (A) despite having finite k^{th} moment for every $k = 0, 1, \dots$. It is natural to expect that Assumption (A) can be relaxed to require that $\mathbf{E}[\sigma^k]$ be finite, for appropriate $k > 2$, in order for Proposition (3.1) to go through. This would still not address the case of long–range dependence, for which $\mathbf{E}[\sigma^2] = \infty$. We present some heuristics below.

Long–range dependence For the Pareto distribution $\mathbf{P}[\sigma > n] = n^{-\alpha}$, $n = 1, 2, \dots$, $1 < \alpha < 2$, not only Assumption (A) fails, but, as mentioned in Section 3, (6) yields infinity. This suggests that $\mathbf{P}_\rho[q > b]$ may not be an analytic function of ρ under long–range dependence. When $c = 1$ simulation work in light traffic does not preclude the possibility that

$\lim_{\rho \rightarrow 0} \rho^{-\alpha} \mathbf{P}_\rho [q > b]$ is the sought after non-trivial limit. On the other hand, the heavy traffic result of Proposition 3.2(b) hints at developing an approximation around the normalized rv $(1 - \rho)^{1/(\alpha-1)} q$. These considerations lead us *to postulate* that, when $c = 1$,

$$\lim_{\rho \rightarrow 0} \rho^{-\alpha} \mathbf{P}_\rho [q > b] = K(b), \quad (21)$$

for some unknown mapping $K : \mathbb{R}_+ \rightarrow \mathbb{R}_+$. Then, taking advantage of Proposition 3.2(b) we propose the approximant

$$\mathbf{P}_\rho [q > b] \approx E_{\alpha-1} \left(-\frac{(\alpha-1)\mathbf{E}[\sigma]}{\Gamma(2-\alpha)} \frac{1-\rho}{\rho^\alpha} b^{\alpha-1} \right), \quad c = 1. \quad (22)$$

This expression is in agreement with the heavy traffic limit (8). In addition, from the Mittag-Leffler function asymptotics given in [5, p. 207], we have

$$E_{\alpha-1}(-x) \sim \frac{1}{x} \frac{1}{\Gamma(2-\alpha)} \quad (x \rightarrow \infty)$$

which ensures that, as $\rho \rightarrow 0$, approximation (22) conforms with the conjectured light traffic limit (21).

Assessing the performance of (22) requires numerical evaluation of the Mittag-Leffler function. In general, calculation based on the series expansion (9) is not recommended. Instead, one can invert the Laplace transform of the Mittag-Leffler law by contour integration along a suitably chosen path in the complex plane. Doing so [15] we arrive at the alternate expression

$$E_\nu(-x) = \frac{\sin(\nu\pi)}{\nu\pi} \int_0^{\pi/2} \frac{e^{-(x \tan \theta)^{1/\nu}}}{1 + \sin(2\theta) \cos(\nu\pi)} d\theta, \quad x \geq 0, \quad 0 < \nu < 1,$$

which is evaluated by numerical integration. We then test approximation (22) for traffic intensities $\rho = 0.2, 0.5$ and 0.8 . Under long-range dependence simulation estimates converge very slowly; moreover confidence intervals based on the regenerative method cannot be constructed, because the underlying period has infinite variance. In the results shown the runs were 10^9 time slots long, and by that time the estimates had stabilized. The log-log scale plots in Figures 2 and 3 correspond to two Pareto distributions with parameters $\alpha = 1.5$ and $\alpha = 1.7$. Observe that the heavier $\alpha = 1.5$ Pareto tail induces larger tail probabilities than $\alpha = 1.7$, at the same traffic intensities. In both Figures 2 and 3 we see that simulated and approximate values are very close, suggesting that expression (22) provides a satisfactory approximation. Note also the almost linear shape of the curves in log-log scale, reflecting the power law asymptotics of the queue size distribution announced in [7, 8, 9].

References

- [1] O. J. Boxma and J. W. Cohen. Heavy traffic analysis for the GI/GI/1 queue with heavy tailed distributions. Technical Report PNA-R9710, CWI, Amsterdam, 1997.

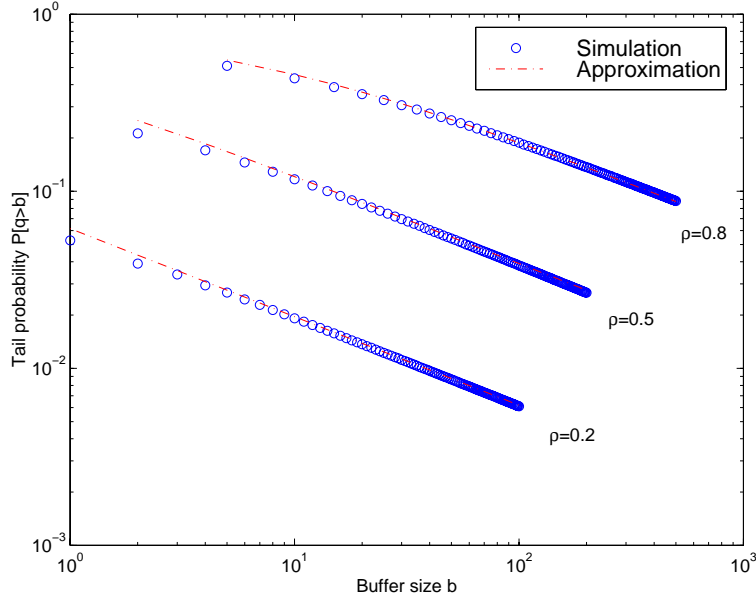


Figure 2: Pareto $\alpha = 1.5$ burst duration.

- [2] A. Brandt, M. Brandt, and H. Sulanke. A single server model for packetwise transmission of messages. *Queueing Systems – Theory and Applications*, 6:287–310, 1990.
- [3] J. W. Cohen. Superimposed renewal processes and storage with gradual input. *Stochastic Processes and their Applications*, 2:31–58, 1974.
- [4] D. R. Cox. Long-range dependence: A review. In H. A. David and H. T. David, editors, *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, Ames (IA), 1984.
- [5] A. Erdélyi. *Higher Transcendental Functions*, volume 3. McGraw-Hill, New York (NY), 1955.
- [6] P. J. Fleming and B. Simon. Interpolation approximations of sojourn time distributions. *Operations Research*, 39(2):251–260, 1991.
- [7] P. R. Jelenković and A. A. Lazar. Multiplexing on–off sources with subexponential on periods: part I. In *Proceedings of IEEE Infocom 97*, Kobe (Japan), April 1997.
- [8] Z. Liu, Ph. Nain, D. Towsley, and Z.-L. Zhang. Asymptotic behavior of a multiplexer fed by a long-range dependent process. *Journal of Applied Probability*. To appear.
- [9] M. Parulekar and A. M. Makowski. $M|G|\infty$ input processes : A versatile class of models for network traffic. In *Proceedings of IEEE Infocom 97*, Kobe (Japan), April 1997.
- [10] M. Parulekar and A. M. Makowski. Tail probabilities for $M|G|\infty$ processes (I): Preliminary asymptotics. *Queueing Systems – Theory and Applications*, 27:271–296, 1997.

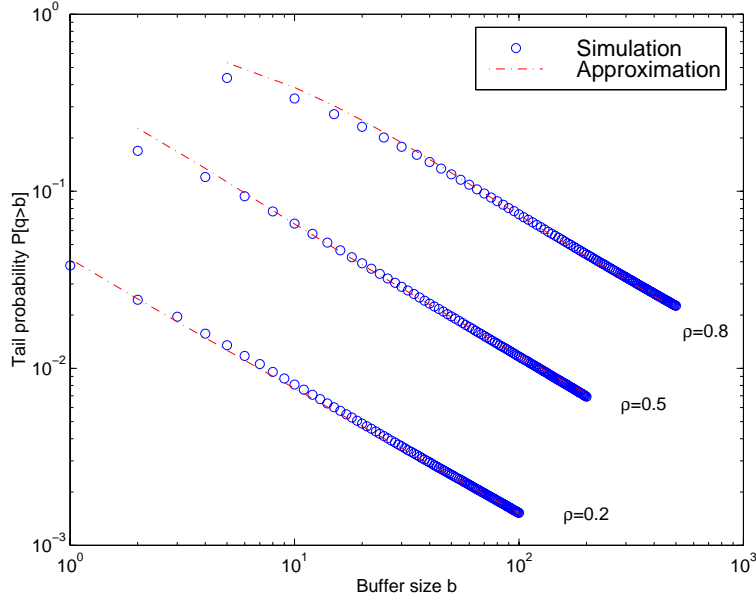


Figure 3: Pareto $\alpha = 1.7$ burst duration.

- [11] M. I. Reiman and B. Simon. An interpolation approximation for queueing systems with Poisson input. *Operations Research*, 36:454–469, 1988.
- [12] M. I. Reiman and B. Simon. Open queueing systems in light traffic. *Mathematics of Operations Research*, 14:26–59, 1989.
- [13] M. Roughan, D. Veitch, and M. Rumsewicz. Computing queue-length distributions for power-law queues. In *Proceedings of IEEE Infocom 98*, San Francisco (CA), April 1998.
- [14] K. Sigman and G. Yamazaki. Fluid models with burst arrivals: A sample path analysis. *Probability in the Engineering and Informational Sciences*, 6:17–27, 1992.
- [15] K. P. Tsoukatos. PhD thesis, University of Maryland, College Park, MD. In preparation.
- [16] K. P. Tsoukatos and A. M. Makowski. Heavy traffic limits associated with $M|G|\infty$ input processes. *Queueing Systems – Theory and Applications*, 1999. To appear.